MS. KAMILLA KJÆRGAARD JENSEN (Orcid ID : 0000-0001-9217-7142)

DR. MASSIMO ANDREATTA (Orcid ID : 0000-0002-8036-2647)

Article type : Original Article

Improved methods for predicting peptide binding affinity to MHC class II molecules

Short title: Improved MHC class II peptide binding predictions

Kamilla Kjærgaard Jensen[1], Massimo Andreatta[2], Paolo Marcatili[1], Søren Buus[3], Jason A. Greenbaum[4], Zhen Yan[4], Alessandro Sette[5,6], Bjoern Peters[5,6], and Morten Nielsen[1,2,*]

[1]Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, 1650 San Martín, Buenos Aires, Argentina

[3]Department of Immunology and Microbiology, Faculty of Health Sciences, University of Copenhagen, Denmark

[4]Bioinformatics Core Facility, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA,
[5]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA

[6]University of California San Diego, Department of Medicine, La Jolla, CA 92037, USA

*corresponding author: mniel@bioinformatics.dtu.dk

Keywords: MHC binding specificity, affinity predictions, peptide-MHC binding, T-cell epitope, immunogenic peptides.

Abbreviations: Major histocompatibility complex (MHC), MHC class II (MHC-II), MHC class I (MHC-I), T-cell receptors (TCRs), Immune Epitope Database (IEDB), peptide flanking regions (PFR), area under the ROC curve (AUC), leave-one-molecule-out (LOMO), unweighted pair group method with arithmetic mean (UPGMA), human leukocyte antigen (HLA), histocompatibility 2 (H-2)

Abstract

Major histocompatibility complex class II (MHC-II) molecules are expressed on the surface of professional antigen presenting cells where they display peptides to T helper cells, which orchestrate the onset and outcome of many host immune responses. Understanding which peptides will be presented by the MHC-II molecule is therefore important for understanding the activation of T helper cells and can be used to identify T-cell epitopes. We here present updated versions of two MHC class II peptide binding affinity prediction methods, NetMHCII and NetMHCIIpan. These were constructed using an extended data set of quantitative MHC-peptide binding affinity data obtained from the Immune Epitope Database covering HLA-DR, HLA-DQ, HLA-DP and H-2 mouse molecules. We show that training with this extended data set improved the performance for peptide binding predictions for both methods. Both methods are publicly available at www.cbs.dtu.dk/services/NetMHCII-2.3 and www.cbs.dtu.dk/services/NetMHCIIpan-3.2.

Introduction

Major histocompatibility complex class II (MHC-II) molecules are found on the surface of antigen presenting cells where they present peptides derived from extracellular proteins to T helper cells (1). Many peptide-MHC complexes are presented on the surface of antigen presenting cells, but only peptides recognized by T-cell receptors (TCRs) will trigger an immune response, and are referred to as T-cell epitopes. Identifying T-cell epitopes is important for the general understanding of cellular immunity and the design of peptide-based diagnostics, therapeutics, and vaccines (2). The MHC-II molecule is a heterodimeric glycoprotein that consists of an α- and a β-chain. In humans, these two chains are encoded in the human leukocyte antigen (HLA) gene complex in one of three loci called: HLA-DR, -DP and -DQ (3). In mice, the MHC-II chains are encoded in the histocompatibility 2 (H-2) locus. Each locus is comprised of many different allelic variants which makes the

MHC-II molecule highly polymorphic (4). Peptides presented by the MHC-II molecule bind to a binding groove formed by residues of the MHC α- and the β-chain. The peptide-binding groove is open at both ends and therefore allows binding of peptides with different lengths (5). Even though the MHC-II molecule can accommodate peptides of variable lengths the most abundant peptides found in nature are between 13 and 25 residues long (6). The part of the peptide ligand that primarily interact with the MHC binding groove is called the peptide binding core and it usually 9 amino acids long (7) with anchor residues at positions P1, P4, P6 and P9 (8). The peptide-MHC binding affinity is primarily determined by the amino acid sequence of the peptide-binding core. However, it has been shown that peptide flanking regions (PFRs) on either side of the binding core affect peptide-MHC binding and, thereby ultimately also influence the peptide immunogenicity (9),(7).

There are therefore many factors that make it difficult to predict peptide binding affinities to MHC-II molecules, including the polymorphic nature of MHC-II molecules, the variations in peptide length, the influence of the peptide flanking regions and the identification of the correct peptide binding core. All these factors complicate the task of predicting peptide binding affinities to MHC-II molecules; most methods therefore still have a low performance compared to MHC class I (MHC-I) peptide binding prediction methods. Earlier work has demonstrated that the prediction performance of both NetMHCII and NetMHCIIpan is dependent on the amount of peptide binding data (10),(11) and one would therefore expect the two methods to improve in performance if retrained on an extended peptide binding data set. We have here investigated if this is indeed the case.

Identifying T-cell epitopes is difficult because of the large diversity in potentially binding peptides. However, as peptide-MHC binding is a prerequisite for T-cell immunogenicity, and multiple studies have shown that there is a strong correlation between MHC peptide binding strength and peptide immunogenicity (12)(13)(14). It is therefore desirable to have accurate and reliable peptide binding affinity prediction methods that can be used for *in silico* screening peptides with the purpose of identifying T-cell epitopes that match MHC-II molecules in a given host. Given this, many different methods have been developed, including NetMHCII (15), NetMHCIIpan (16), TEPITOPE (17), TEPITOPEpan (18),PROPRED (19), RANKPEP (20)(21) and SVRMHC (22). Both NetMHCII (15) and NetMHCIIpan (16) have been shown to be among the best methods for predicting binding affinities to MHC class II molecules (2),(8),(23). These two methods are trained using the NNAlign framework (15),(24),(25)

and are based on ensembles of artificial neural networks which are trained on quantitative peptide binding affinity data from IEDB (26). One of the main differences between NetMHCII and NetMHCIIpan is that NetMHCII is a collection of individual networks for each MHC molecule whereas NetMHCIIpan contains a single universal network that can predict peptide binding affinities for all MHC molecules of known protein sequence.

NetMHCII and NetMHCIIpan predict peptide binding affinities to MHC class II molecules covering HLA-DR, HLA-DQ, HLA-DP and H-2 mouse molecules. The main difference between the two methods is that NetMHCII only predict peptide binding affinities to MHC molecules for which it has been trained, while NetMCHIIpan can predict peptide binding affinities to any MHC molecule with a known protein sequence. As mentioned above there is a strong correlation between MHC binding strength and peptide immunogenicity and the two methods have been used extensively as a guide to identify T-cell epitopes which can be used in the design of peptide-based diagnostics, therapeutics, and vaccines.

In this paper, we present updated versions of our binding affinity prediction methods, NetMHCII and NetMHCIIpan, trained on an extended data set of more than 100,000 quantitative peptide-binding measurements from IEDB (26), covering 36 HLA-DR, 27 HLA-DQ, 9 HLA-DP, as well as 8 mouse MHC-II molecules. We then evaluate the performance of these new versions using a set of large scale benchmarks to investigate how the extended data set improves the predictive performance of the two both methods.

Materials and methods

Data sets

The data set used to generate the new versions of NetMHCII and NetMHCIIpan contains peptide-MHC II binding affinities retrieved from the Immune Epitope Database (IEDB, www.iedb.org) in 2016. All data points are experimental IC50 binding values which were log-transformed to fall in the range between 0 and 1 using the relation 1−log(IC50nM)/log(50,000) as explained in (27). This 2016 data set contains 134,281 data points, covering 36 HLA-DR, 27 HLA-DQ, 9 HLA-DP and 8 H-2 molecules. The data set was split into five groups by clustering the common motif of peptides as described by Nielsen et al. (28) and these five groups were used

for a five-fold cross validation. This 2016 data set is publicly available at www.cbs.dtu.dk/suppl/immunology/NetMHCIIpan-3.2. The data set used to develop the previous versions of NetMHCII and NetMHCIIpan is available at www.cbs.dtu.dk/suppl/immunology/NetMHCIIpan-3.0.

A summary of the data included in the 2013 and 2016 data sets is shown in Table 1 and a description of the full 2016 data set is available in supplementary table 1.

Network training

The NetMHCII method was implemented as described by Nielsen and Lund (15) and the NetMHCIIpan method was implemented as described by Andreatta et al. (16). NetMHCII is an allele-specific method that contains a specific predictor for each MHC molecule in the data set and it can therefore only predict binding affinities for MHC molecules found in the training data, whereas NetMHCIIpan is a pan-specific method that can make predictions for any MHC molecule of known protein sequence. To achieve its pan-specificity, NetMHCIIpan incorporates information about the MHC-II molecule, using a pseudo sequence consisting of residues which are considered important for peptide binding. This pseudo sequence is constructed using the method described by Karosiene et al. (11) and is composed of 34 residues, 15 residues from the α-chain and 19 residues from the β-chain. Both methods were trained using a five-fold cross-validation setup. For each fold, we generate a network ensemble of individual networks trained without early stopping for 500 cycles with 10, 15, 40 and 60 hidden neurons using 10 different initial configurations, generating a total of 40 networks. This was done for each of the 5 training/test set combination leading to a total of 200 networks. The peptide and the MHC pseudo sequence was encoded using the BLOSUM50 matrix and the peptide flanking regions (PFR) was encoded using the average BLOSUM scores on a maximum window of three amino acids at either end of the binding core (29). For each peptide core, the input to the neural network thus consisted of the peptide core ($9 \times 20 = 180$ inputs), the PFRs ($2 \times 20 = 40$ inputs), the peptide length (2 inputs), the length of the C- and N- terminal PFR's ($2 \times 2 = 4$ inputs), resulting in a total of 226 input values for NetMHCII and 906 for NetMHCIIpan (an additional $34 \times 20 = 680$ input values from the pseudo sequence).

Binding core predictions

To improve the binding core predictions, we include the offset correction step to both NetMHCII and NetMHCIIpan as described by Andreatta et al. (16) and we evaluated the performance of this offset correction using the benchmark data set of 51 crystal structures of peptide-MHC class II complexes as described in (16).

Performance measures

The predictive performance of the different methods was measured using the area under the ROC curve (AUC). To classify peptides into binders and non-binders, a binding threshold of 500 nM was used, classifying all peptides with a IC50 binding value below 500 nM as binders. All performance values shown in this paper are averages of the performance per MHC molecule using only molecules with more than 20 peptides and at least 4 binders.

Leave-one-molecule-out network training

In order to assess the predictive performance of NetMHCIIpan in the situation where a molecule is not part of the training data, a leave-one-molecule-out (LOMO) approach was applied.

To estimate LOMO performance for MHC molecule X, the NetMHCIIpan networks were trained using the five-fold cross-validation setup from above. In the LOMO cross-validation setup all binding data from molecule X were removed from the training sets and all test set only include binding data from molecule X. This setup ensures that the method is trained without peptides binding to molecule X and it can therefore be used to evaluate the ability of the method to predict peptide binding of uncharacterized MHC-II molecules.

Nearest neighbor distance calculation

The nearest neighbor distance is estimated from the alignment score of the HLA pseudo sequences using the relation $= \frac{s(A,B)}{\sqrt{s(A,A)\cdot s(B,B)}}$. In this equation s(A,B) is the BLOSUM50 alignment score between the pseudo

sequences for MHC molecule A and B, respectively (29). Nearest neighbors are found from the subset of molecules characterized with at least 50 data points and at least 10 binders.

Sequence logos

Sequence logos were constructed from the predicted binding cores of the top 1% strongest predicted binders using 200,000 natural random 15-mer peptides and was visualized using Seq2Logo (30) with default settings.

Generation of HLA-II distance trees

The HLA-II distance tree was generated for each of the HLA-DR, -DQ and -DP molecule in our data set using MHCCluster (31). To make the tree we first predicted the binding affinity for 200,000 natural random 15-mer peptides using the new version of NetMHCIIpan. We then used MHCCluster to find the functional similarity between any two MHC molecules. MHCCluster calculates the similarity between two MHC molecules by correlating the union of the predicted top 10 % strongest binding peptides. Using the bootstrap method in MHCCluster we generated 100 distance matrices and converted these to distance trees using the unweighted pair group method with arithmetic mean (UPGMA) clustering. These trees were then combined into a consensus tree and visualized in SplitsTree (32). Sequence logos were constructed as explained above.

T-cell epitope benchmark

A set of MHC-II restricted T-cell epitopes identified by multimer/tetramer staining assays was downloaded from IEDB. Only fully typed restrictions were included; that is, fully typed α- and β-chains for HLA-DQ and HLA-DP, and a fully typed β-chain for HLA-DR (where the alpha chain is invariant). Epitopes with non-natural amino acids were excluded. Also, epitopes with identical match to the peptides in the training data were excluded. The source protein sequence for each epitope was identified by mapping the annotated IEDB protein ID to the NCBI protein database. The final validation data set consisted of 1698 epitopes, restricted to 33 distinct MHC class II molecules. For performance evaluation, the epitope source protein was split into overlapping peptides of the length of the epitope, and AUC and Frank values were calculated for each epitope-MHC pair annotating

the epitopes as positive and all other as negatives. Here, Frank is the ratio of the number of peptides with a prediction score higher than the positive peptide to the number of peptides contained within the source protein. Hence, the Frank value is 0 if the positive peptide has the highest prediction value of all peptides within the source protein and a value of 0.5 in cases in which an equal number of peptides has a higher and lower prediction value compared with the positive peptide.

Results

Comparing NetMHCII and NetMHCIIpan on a shared evaluation set

Using the data set from 2016, we retrained NetMHCII (15) and NetMHCIIpan (11) using a five-fold cross-validation setup to generate two new versions of these methods, named NetMHCII-2.3 and NetMHCIIpan-3.2. We then investigated how these new versions performed compared to the previous versions, which are NetMHCII-2.2 and NetMHCIIpan-3.1, trained on the 2013 data set. To make the comparison, we used the same five-fold cross-validation setup and compared peptide data points in common between the 2013 and 2016 data sets. The result from this analysis in shown in Table 2.

The new versions of NetMHCII and NetMHCIIpan improved performance compared to the older versions (table 2). This performance gain is however not statistically significant ($p>0.1$ in both cases). Another interesting point is that the allele-specific NetMHCII-2.3 obtains a higher average performance than the pan-specific NetMHCIIpan-3.2. This effect will be discussed later.

Performance of NetMHCIIpan on new data points for common MHC molecules

Using the five-fold cross-validation setup, we then evaluated the performance of the two versions of NetMHCII and NetMHCIIpan using only the subset of new peptides for the MHC molecules common between the old and the new data sets. The result of this analysis is shown in table 3 and it demonstrates a significant gain in predictive performance of the new versions (NetMHCII, p-value < 0.005 and NetMHCIIpan, p-value < 0.001, using paired t-test). This result thus underlines the importance of expanding the size of the training data even for previously characterized MHC molecules.

Binding core predictions

We evaluated the accuracy for binding core identification of the two updated MHC class II binding prediction methods on the data set of peptide-MHC crystal structures described by Andreatta et al. (16). Overall we find that i) the inclusion of the offset correction described earlier to align the individual network in the network ensemble has a substantial impact on the accuracy of binding core identification for both methods, and ii) that the overall accuracy of both methods is improved compared to the earlier version. For details see supplementary table 2.

Performance of a consensus method

For predicting binding affinities to MHC class I, it has been shown that a simple combination of the predictions from NetMHC (27) and NetMHCpan (10) gives a higher performance than using each method individually (33). We therefore made a similar combination of the predictions from NetMHCII-2.3 and NetMHCIIpan-3.2 to investigate if the performance could be improved for MHC class II using this consensus approach. In the consensus method, we use an average of the prediction scores (values between 0 and 1) from NetMHCII-2.3 and NetMHCIIpan-3.2 to define the consensus method. The result of this analysis is shown in figure 1 and detailed performance values are found in supplementary table 3. Figure 1A shows that the combination of NetMHCII-2.3 and NetMCHIIpan-3.2 has a significantly improved performance compared to each individual method and figure 1B shows that NetMHCIIpan-3.2 outperforms NetMHCII-2.3 especially for MHC molecules where only few peptides are found in the data set.

Performance of NetMHCIIpan for previously uncharacterized MHC molecules

For NetMHCIIpan, we also tested the performance on MHC molecules that were not part of the 2013 data set (see table 4). As expected, we observed that the new version of NetMHCIIpan had a significant increase in the predictive performance when compared to the previous version of NetMHCIIpan (p-value = $3.6 \cdot 10^{-5}$, using a

paired t-test); this result thus demonstrates the importance of expanding the allotypic coverage of the training data.

Leave-one-molecule-out performance

The pan-specific method is capable of making predictions for uncharacterized MHC molecules, so to assess the predictive performance of the NetMHCIIpan method in these situations we conducted a leave-one-molecule-out (LOMO) experiment. In the LOMO, the binding data for the MHC molecule in question were excluded from training and the resulting model were then evaluated using only binding data for the MHC molecule in question (for details see Materials and Methods). The LOMO experiment was made for all MHC molecules shared between the 2013 and the 2016 data sets, and the performance evaluated on peptides shared between the two data sets. The result of this LOMO benchmark is shown in table 5, together with the pseudo distances of the MHC molecule to each of the two training data sets estimated from the nearest neighbor sequence similarity as described in Materials and Methods.

Table 5 shows an increased performance for NetMHCIIpan-3.2-LOMO compared to netMHCIIpan-3.1-LOMO. This gain is in general most pronounced for the MHC molecules that share a decrease in the pseudo sequence distance.

To further investigate this last observation, the LOMO performance evaluation was extended to include all MHC molecules in the 2016 data set. The result from this analysis is shown in figure 2 with a scatterplot of the relationship between the distance to the nearest neighbor in the training data set and the LOMO performance. The complete data used to create figure 2 can be found in supplementary table 4. The figure shows that the HLA-DQ and the HLA-DP molecules have close nearest neighbors while the HLA-DR and H-2 molecules tend to have more distant neighbors. This figure also demonstrates a weak but statistically significant (p-value of 0.04 with exact permutation test) correlation between the LOMO performance and the distance to the nearest neighbor in the training data. This is in agreement with earlier findings for both MHC-I and MHC-II molecules (10)(11) and shows how the predictive performance of the pan-specific method depends on the distance to the nearest neighbor.

Distance tree for HLA molecules

Having arrived at the final retrained versions of NetMHCIIpan, we next use the MHCCluster method (31) to evaluate the similarities of binding motives between the HLA molecules included in the 2016 training data. In short, the MHCCluster method estimates the similarity between two MHC molecules using the correlation between predicted binding values for a large set of random natural peptides. The similarity is 1 if the two molecules have a perfect binding specificity overlap and −1 if the two molecules share no specificity overlap (for details see Materials and Methods). Comparing the binding pattern similarity between any two HLA class II molecule in the 2016 training data, we constructed the distance tree shown in figure 3. This figure confirms the earlier findings by Karosiene et al. (11), i) the different loci shows limited overlap in binding preference, ii) HLA-DP is less diverse compared to HLA-DQ and HLA-DR, and iii) the diversity of HLA-DQ can largely be split into 3 groups; one with preference for negatively charged amino acids toward to the C-terminal, one with a preference for positively charged amino acids towards the C-terminus, and one with preference for small amino acids at the anchor positions.

T-cell epitope benchmark

We next evaluated the predictive performance of the two NetMHCIIpan methods on an IEDB T-cell epitope data set. We queried the IEDB for MHC-II restricted epitopes identified by tetramer/multi-mer staining, which is the gold-standard for epitope identification with known MHC restriction. For each epitope-MHC class II pair, we calculated AUC and Frank values for the two NetMHCIIpan methods by predicting binding affinities to the MHC class II restriction element of the epitope for all overlapping peptides with the same length as the epitope in the source protein sequence, annotating the epitope as positive and the remaining peptides as negative. This annotation is very stringent since peptides that share the same ligand binding-core are counted as negatives even though they could be presented by the human MHC molecule; the setup will therefore most likely underestimate the predictive performance. The details from this analysis are found in supplementary table 5 and the results are summarized in figure 4.

Recall that the Frank value is 0 if the positive peptide has the highest prediction value of all peptides within the source protein, and a value of 0.5 in cases where an equal number of peptides has a higher and lower prediction value compared to the positive peptide. Figure 4 A shows that the Frank score for NetMHCIIpan-3.1 is significant lower than NetMHCIIpan-3.1. It further shows that NetMHCIIpan-3.2 has a median below 0.2 indicating that the positive peptide was found among the top 20% of the peptides from the source protein if sorted on their predicted peptide binding affinity. Figure 4 B demonstrates a significant improvement in the AUC performance of NetMHCIIpan-3.2 compared to NetMHCIIpan-3.1. We speculate that the gain in predictive performance of NetMHCIIpan-3.2 could be attributed to at least two factors, the inclusion of binding data for additional MHC class II molecules in the training data, and the expansion of the number of data points for MHC class II molecules already included in the old training data. Figures 4 C and D quantify that both of these factors indeed contribute to the performance gain. Figure 4 C shows the performance gain as a function of the change in distance of the query molecule to the nearest neighbor of the training data. From this plot, we see that the gain in predictive performance is related to a decrease in the nearest neighbor distance, and hence directly related to the inclusion of binding data for additional MHC class II molecules in the new data set. Figure 4 D shows the performance gain as a function of the change in the number of data points between the two data sets used for training. We here only include molecules shared between the two data sets used for training NetMHCIIpan-3.1 and NetMHCIIpan-3.2, as we in the previous analysis demonstrated how the distance to the nearest neighbor influences the performance. Figure 4 D shows that the gain in performance is correlated to change in the number of data points for the given MHC molecules. This indicates that the performance gain of the new NetMHCIIpan version is also driven by the increase in the number of data points for molecules already included in the 2013 data set. The one data point in Figure 4 C and D with increased nearest neighbor distance and decreased number of data points corresponds to the HLA-DPA10103-DPB10201 molecule for which faulty data was removed in the 2016 dataset.

Discussion

The genomic region encoding the MHC-II molecule is extremely polymorphic comprising several thousand alleles and it is therefore difficult to produce enough experimental data to characterize the peptide binding

preference for all existing MHC-II molecules. Because of this, most MHC class II molecules are still only represented with very few or no binding data, limiting the coverage and performance of previous binding affinity prediction methods. We have therefore updated our two binding affinity prediction methods, NetMHCII and NetMHCIIpan using updated and extended data sets. For several large-scale benchmarks, this improved the predictive performance for both methods.

Comparing NetMHCII and NetMHCIIpan

Using the data points shared by the old and the updated data sets, we first compared the different versions of NetMHCII and NetMHCIIpan. We showed how the new versions of the methods outperformed the previous versions both for NetMHCII and NetMHCIIpan. We then evaluated the performance of the two versions of the methods using only "new" peptides, for the MHC molecules covered both by the old and the updated data sets. The result of this analysis showed that both methods on this data set gained a significant improvement in the predictive performance, thus supporting the importance of expanding the size of the training data even for MHC molecules already characterized by binding data. When evaluating new peptides one has to keep in mind that MHC binding predictors are often used to select peptides for experimental validation and new data sets may be less diverse than historic data sets generated sampling the entire space of a given set of protein sequences (34).

The main difference between NetMHCII and NetMHCIIpan is that NetMHCII is an allele-specific method trained separately for each MHC molecule, whereas NetMHCIIpan is a pan-specific method that contains a single ensemble of networks using information from all MHC molecules in the data set. We would therefore expect that the allele-specific method outperforms the pan-specific method for MHC molecules where sufficient data is available to accurately characterize the binding motif, and we would expect the pan-specific method to outperform the allele-specific method when data is scarcer. This is exactly what we observed when we compared the predictive performance of NetMHCII-2.3 and NetMHCpan-3.2. Earlier work has shown a similar result, namely that when allele-specific neural network prediction algorithms rely on a sufficient number of peptide binders they achieve high predictive performances (33),(35). This illustrates how the allele-specific method is preferable only if a large amount of data is available for the MHC molecule in question, while

highlighting the strength of the pan-specific methods, which can benefit from data of related MHC molecules to make reliable predictions for MHC molecules with limited data. Because of this difference between the allele-specific and the pan-specific method, we implemented a simple combination of two methods as this has been shown to improve the predictive performance for MHC class I molecules (33). This analysis showed that NetMHCIIpan-3.2 outperforms NetMHCII-2.3 for MHC molecules which have been trained with very few peptides, but that a combination of the predictions from the two MHC class II methods still outperformed each individual method.

Leave-one-molecule-out performance for NetMHCIIpan

One of the main powers of the NetMHCIIpan method is that it can predict binding affinities for uncharacterized MHC molecules. To assess the performance of the method in such a task, we constructed a LOMO experiment where we tested the performance of the NetMHCIIpan method for predicting binding affinity for MHC molecules not included in the training data for the method. From this analysis, we could show that the pan-specific method is capable of prediction binding affinity for MHC molecules where no binding affinity data is available and further demonstrate that the predictive performance is dependent on the distance to the nearest neighbor. This last observation indicated that the predictive performance of the NetMHCIIpan method could be further improved by including more uncharacterized MHC molecules into the training data and it is therefore important to generate experimental peptide binding affinity data points in a targeted fashion for MHC molecules not yet characterized.

Distance tree for HLA class II molecules

To understand the different groups of HLA class II molecules, we generated a fictional distance tree using NetMHCIIpan-3.2. The groups shown in this distance tree can be used to understand how peptides interact with different MHC molecules and can be used to discriminate between binders and non-binders. The distance tree can also be used to identify T-cell epitopes with similar properties important for the design of epitope-based vaccines. Another aspect that can be observed for the tree is that most MHC molecules have strong

anchor positions at P1, P4, P6 and P9 which have also been observed in previous studies (8).

T-cell epitope benchmark

Accurate predictions of peptide binding affinities to MHC molecules are important for understanding the cell mediated immune response and for generating better screening methods for cost-effective identification of immunogenic peptides. We therefore wanted to test the predictive performance of the two versions of NetMHCIIpan on a T-cell epitope data set, and doing this we demonstrated how the new version of NetMHCIIpan obtained a significantly improved predictive performance compared to the earlier version. Two main factors explain this performance gain: i) including data for new MHC-II molecules decrease the distance to the nearest neighbor, ii) including an increased number of data points allow the method better characterizing the specificity of a given MHC-II molecule.

In conclusion, we believe that NetMHCII and NetMHCIIpan can be used to improve MHC-II binding predictions and reduce experimental costs for immunologists working within the field of epitope-based vaccine design, and to improve our knowledge about the peptide-MHC interaction, a key event in the cellular immune response.

Acknowledgements

This work was supported by Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200010C.

Disclosures

The authors declare having no competing interests.

References:

1. Castellino F, Zhong G, Germain RN. Antigen presentation by MHC class II molecules: Invariant chain function, protein trafficking, and the molecular basis of diverse determinant capture. Hum Immunol. 1997;54(2):159–69.

2. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics. 2008;9(Suppl. 12):S22.

3. Traherne JA. Human MHC architecture and evolution: Implications for disease association studies. Int J Immunogenet. 2008;35(3):179–92.

4. Nielsen M, Lund O, Buus S, Lundegaard C. MHC Class II epitope predictive algorithms. Immunology. 2010;130(3):319–28.

5. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, et al. Three-Dimensional Structure of the Human Class II Histocompatibility Antigen HLA-DR1. J Immunol. 2015;194(1):5–11.

6. Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DAA, et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature. 1992;358(6389):764–8.

7. Holland CJ, Cole DK, Godkin A. Re-directing CD4 + T cell responses with the flanking residues of MHC class II-bound peptides: The core is not enough. Front Immunol. 2013;4:Article 172.

8. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. Brief Bioinform. 2012;13(3):350–64.

9. Arnold PY, La Gruta NL, Miller T, Vignali KM, Adams PS, Woodland DL, et al. The Majority of Immunogenic Epitopes Generate CD4+ T Cells That Are Dependent on MHC Class II-Bound Peptide-Flanking Residues. J Immunol. 2002;169(2):739–49.

10. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007;2(8).

11. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013;65(10):711–24.

12. Iwai LK, Yoshida M, Sidney J, Shikanai-Yasuda MA, Goldberg AC, Juliano MA, et al. In silico prediction of peptides binding to multiple HLA-DR molecules accurately identifies immunodominant epitopes from gp43 of Paracoccidioides brasiliensis frequently recognized in primary peripheral blood mononuclear cell responses from sensitized individuals. Mol Med. 2003;9(9):209–19.

13. Mustafa AS, Shaban FA. ProPred analysis and experimental evaluation of promiscuous T-cell epitopes of three major secreted antigens of Mycobacterium tuberculosis. Tuberculosis. 2006;86(2):115–24.

14. Al-Attiyah R, Mustafa AS. Computer-Assisted Prediction of HLA-DR Binding and Experimental Analysis for Human Promiscuous Th1-Cell Peptides in the 24 kDa Secreted Lipoprotein (LppX) of Mycobacterium tuberculosis. Scand J Immunol. 2004;59(1):16–24.

15. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009;10(1):1471.

16. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics. 2015;67(11–12):641–50.

17. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999;17(6):555–61.

18. Zhang L, Chen Y, Wong H, Zhou S, Mamitsuka H. TEPITOPEpan : Extending TEPITOPE for Peptide Binding Prediction Covering over 700 HLA-DR Molecules. PLoS One. 2012;7(2):e30483.

19. Singh H, Raghava GPS. ProPred: Prediction of HLA-DR binding sites. Bioinformatics. 2002;17(12):1236–7.

20. Reche PA, Glutting JP, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. Hum

Immunol. 2002;63(9):701–9.

21. Reche PA, Glutting JP, Zhang H, Reinherz EL. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics. 2004;56(6):405–19.

22. Wan J, Liu W, Xu Q, Ren Y, Flower DR, Li T. SVRMHC prediction server for MHC-binding peptides. BMC Bioinformatics. 2006;7(1):463.

23. Sette A, Peters B, Wang P, Sidney J, Dow C, Mothe B. A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach. Plos Comput Biol. 2008;4(4):e1000048.

24. Nielsen M, Andreatta M. NNAlign : a platform to construct and evaluate artificial neural network models of receptor – ligand interactions. Nucleic Acids Res. 2017;45(1):344–9.

25. Andreatta M, Schafer-nielsen C, Lund O, Buus S, Nielsen M. NNAlign : A Web-Based Prediction Method Allowing Non- Expert End-User Discovery of Sequence Motifs in Quantitative Peptide Data. PLoS One. 2011;6(11):e26781.

26. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark D, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43(D1):D405–12.

27. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12(5):1007–17.

28. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007;8(1):238.

29. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, et al. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. PLoS Comput Biol. 2008;4(7):e1000107.

30. Thomsen MCF, Nielsen M. Seq2Logo: A method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided

representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;40(W1):281–7.

31. Thomsen M, Lundegaard C, Nielsen M. MHCcluster, a method for functional clustering of MHC molecules. Immunogenetics. 2013;65(9):655–65.

32. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006;23(2):254–67.

33. Karosiene E, Lundegaard C, Lund O. NetMHCcons : a consensus method for the major histocompatibility complex class I predictions. 2012;177–86.

34. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC Bioinformatics. 2014;15(1):241.

35. Zhang H, Lundegaard C, Nielsen M. Pan-specific MHC class I predictors : a benchmark of HLA class I pan-specific prediction methods. Bioinformatics. 2009;25(1):83–9.

Figure legends:

Figure 1: Performance of NetMHCII-2.3 and NetMHCIIpan-3.2 together with the combination method. (A) The average performance per MHC molecule of NetMHCII-2.3, NetMHCIIpan-3.2 and the combination method, including the significance between the methods. P-values where found using a paired T-test using the predictions per molecule found in supplementary table 3. (B) The average predictive performance of the MHC molecules in the data set as a function of the number of peptides.

Figure 2: Predictive performance for NetMHCIIpan-3.2 LOMO on the MHC class II molecules from the 2016 data set as a function of distance to the nearest neighbor. Each HLA II isotype and H-2 molecules are displayed in different colors and the dashed line represents the least square fit for the data.

Figure 3: Distance tree for all HLA molecules found in our data set generated using the MHCCluster method. Sequence logos shows the motif of the predicted binding core for each HLA and were generated using Seq2Logo (30). The names are colored according to the type of HLA, HLA-DR in red, HLA-DQ in blue and HLA-DP in green.

Figure 4: Performance of NetMHCIIpan-3.1 and NetMHCIIpan-3.2 using the T-cell epitope benchmark set. (A) The average Frank performance per MHC molecule for the two versions of NetMHCIIpan. (B) The average AUC performance per MHC molecule for the two versions of NetMHCIIpan. (C) The change in the distance to the nearest neighbor between the two data sets used for training the old and the new version of NetMHCIIpan as a function of the change in distance to the nearest neighbor. (D) the change in the number of data points between the two data sets used for training NetMHCIIpan-3.1 and NetMHCIIpan-3.2 as a function of the change in the performance, including only MHC molecules where the pseudo sequence did not change between two data sets. The dashed line in the two scatterplots represents the least square fit for the data.

Table 1: Description of the two MHC class II peptide binding data sets.

Table 2: Comparing prediction from the old and the new versions of NetMHCII and NetMHCIIpan trained using a five-fold cross-validation on the set of data points common between the two data sets. For each MHC molecule, we show the total number of peptides, the number of binders, the AUC performance. The different methods included are the NetMHCII and NetMHCIIpan methods training on the original 2013 data set (versions 2.2 and 3.1), and the versions of the two methods trained on the extended 2016 data set (versions 2.3 and 3.2). The highest performance for NetMHCII and NetMHCIIpan is highlighted in bold.

Table 3: Comparing predictions from the old (versions 2.2 and 3.1), and the new version (versions 2.3 and 3.2), of NetMHCII and NetMHCpan using the 5-fold cross-validation setup and evaluating on the subset of new peptides using only MHC molecules shared between the 2013 and 2016 data sets. For each MHC molecule, we show the total number of peptides, the number of binders and the AUC performance for the different versions. Highlighted in bold is the highest performance between the two NetMHCII and NetMHCIIpan methods.

Table 4: Comparing prediction from the old and the new version of NetMHCIIpan using the 5-fold cross-validation setup on the set of MHC molecules found in the 2016 data set but not in the 2013 data set. For each molecule, we show the total number of peptides, the number of binders and the AUC performance for the two NetMHCIIpan versions. In bold is highlighted the highest performance of the two versions 3.1 and 3.2 of NetMHCIIpan. Highlighted in bold is the highest performance between the two methods.

Table 5: Comparing leave-one-molecule-out predictions from the old and the new method on the set of data points common between the two data sets. For each molecule, we show the number of peptides, the number of binders, the AUC performance for the old (3.1) and new (3.2) methods, and the distance to the nearest neighbor for the old and new data set. Nearest neighbors are found from the subset of molecules in the training data characterized with at least 50 data points and at least 10 binders. Highlighted in bold is the highest performance between the two methods.

	Data set 2013	**Data set 2016**
# data points	52062	134281
Type of alleles	24 HLA-DR	36 HLA-DR
	6 HLA-DQ	27 HLA-DQ
	5 HLA-DP	9 HLA-DP
	2 H-2	8 H-2

Table 1

Molecule	#peptides	#binders	NetMHCII-2.2	NetMHCII-2.3	NetMHCIIpan-3.1	NetMHCIIpan-3.2
DRB1_0101	2754	2635	0.817	**0.822**	0.828	**0.830**
DRB1_0301	1403	379	**0.832**	0.826	0.829	**0.835**
DRB1_0401	1639	695	**0.801**	0.791	**0.804**	0.798
DRB1_0404	542	331	**0.783**	0.768	**0.813**	0.810
DRB1_0405	1438	595	**0.862**	0.860	**0.852**	0.844
DRB1_0701	1619	806	**0.858**	0.857	0.852	**0.857**
DRB1_0802	1310	400	0.757	**0.767**	**0.753**	0.749
DRB1_0901	841	560	0.746	**0.761**	0.777	**0.779**
DRB1_1101	1604	730	0.876	0.876	0.875	**0.876**
DRB1_1302	1351	463	0.811	**0.823**	0.801	**0.810**
DRB1_1501	1601	672	0.818	**0.820**	0.817	**0.831**
DRB3_0101	1266	267	0.835	**0.846**	**0.835**	0.824
DRB4_0101	1329	467	0.840	**0.841**	**0.832**	0.817
DRB5_0101	1606	765	**0.852**	0.847	**0.855**	0.846
H-2-IAb	525	125	0.850	**0.857**	0.849	**0.868**
H-2-IAd	100	24	0.718	**0.809**	0.734	**0.808**
HLA-DPA10103-DPB10401	1075	458	0.957	**0.960**	0.956	**0.961**
HLA-DPA10201-DPB10101	1180	558	0.949	0.949	**0.949**	0.948
HLA-DPA10201-DPB10501	1114	415	**0.957**	0.954	**0.949**	0.948
HLA-DPA10301-DPB10402	1193	498	**0.958**	0.957	**0.957**	0.952
HLA-DQA10101-DQB10501	990	246	0.856	**0.890**	0.834	**0.857**
HLA-DQA10102-DQB10602	1121	503	0.838	**0.901**	0.877	**0.887**
HLA-DQA10301-DQB10302	1461	330	**0.824**	0.820	**0.796**	0.774
HLA-DQA10401-DQB10402	1436	516	0.919	**0.923**	**0.915**	0.903
HLA-DQA10501-DQB10201	1386	477	0.898	**0.901**	**0.886**	0.883
HLA-DQA10501-DQB10301	1274	530	**0.893**	0.873	**0.881**	0.860
Average			**0.856**	**0.863**	**0.856**	**0.858**

Table 2

Molecule	#peptides	#binders	NetMHCII-2.2	NetMHCII-2.3	NetMHCIIpan-3.1	NetMHCIIpan-3.2
DRB1_0101	7909	3975	**0.855**	0.816	**0.839**	0.824
DRB1_0301	4086	1120	0.805	**0.813**	0.783	**0.812**
DRB1_0401	4849	2391	0.780	**0.799**	0.776	**0.812**
DRB1_0404	3169	1549	0.715	**0.787**	0.763	**0.810**
DRB1_0405	2663	1120	0.809	**0.832**	0.817	**0.819**
DRB1_0701	4862	2727	0.828	**0.882**	0.829	**0.880**
DRB1_0802	3273	1669	0.804	**0.845**	0.829	**0.854**
DRB1_0901	3578	1662	**0.845**	0.843	0.835	**0.839**
DRB1_1101	4610	2018	0.833	**0.864**	0.825	**0.862**
DRB1_1302	3259	1824	0.858	**0.906**	0.864	**0.905**
DRB1_1501	3392	1497	0.814	**0.840**	0.821	**0.836**
DRB3_0101	3497	1177	0.899	**0.911**	0.898	**0.904**
DRB4_0101	2764	1121	0.804	**0.836**	0.810	**0.823**
DRB5_0101	3681	1738	0.841	**0.849**	0.843	**0.849**
H-2-IAb	1364	306	**0.942**	0.900	**0.926**	0.909
H-2-IAd	683	297	0.767	**0.820**	0.803	**0.821**
HLA-DPA10103-DPB10201	784	140	**0.968**	0.909	**0.954**	0.917
HLA-DPA10103-DPB10401	1804	328	0.897	**0.907**	0.895	**0.906**
HLA-DPA10201-DPB10101	1382	301	0.838	**0.847**	0.846	**0.861**
HLA-DPA10201-DPB10501	1537	298	0.869	**0.872**	0.839	**0.873**
HLA-DPA10301-DPB10402	1591	423	0.858	0.858	0.859	**0.861**
HLA-DQA10101-DQB10501	2122	569	**0.937**	0.935	**0.928**	0.926
HLA-DQA10102-DQB10602	1710	753	0.868	**0.919**	0.890	**0.910**
HLA-DQA10301-DQB10302	1790	238	0.863	**0.878**	**0.851**	0.843
HLA-DQA10401-DQB10402	1620	412	0.810	**0.876**	0.810	**0.875**
HLA-DQA10501-DQB10201	1656	397	0.849	**0.884**	0.851	**0.881**
HLA-DQA10501-DQB10301	2379	1282	0.915	**0.947**	0.926	**0.946**
Average			**0.847**	**0.866**	**0.849**	**0.865**

Table 3

Molecule	#peptides	#binders	NetMHCIIpan-3.1	NetMHCIIpan-3.2
DRB1_0103	42	4	0.664	**0.678**
DRB1_0402	53	19	0.680	**0.701**
DRB1_0403	59	14	0.767	**0.841**
DRB1_0801	937	390	0.839	**0.844**
DRB1_1001	2066	1521	0.907	**0.923**
DRB1_1104	27	5	0.682	**0.791**
DRB1_1301	1034	520	0.727	**0.857**
DRB1_1502	23	7	**0.688**	0.652
DRB1_1602	1699	989	0.827	**0.883**
DRB3_0202	3334	1055	0.789	**0.869**
DRB4_0103	846	525	0.786	**0.841**
H-2-IAk	115	4	0.426	**0.635**
H-2-IAs	190	48	0.438	**0.825**
H-2-IAu	56	22	**0.790**	0.765
H-2-IEd	245	28	0.623	**0.754**
H-2-IEk	68	40	**0.881**	0.853
HLA-DPA10103-DPB10301	1563	575	0.588	**0.902**
HLA-DPA10103-DPB10402	45	9	**0.815**	0.710
HLA-DPA10103-DPB10601	584	282	**0.996**	0.995
HLA-DPA10201-DPB11401	2302	849	0.696	**0.930**
HLA-DQA10102-DQB10501	833	458	0.606	**0.839**
HLA-DQA10102-DQB10502	800	158	0.825	**0.835**
HLA-DQA10103-DQB10603	462	90	0.802	**0.861**
HLA-DQA10104-DQB10503	883	105	0.787	**0.805**
HLA-DQA10201-DQB10202	944	119	0.779	**0.814**
HLA-DQA10201-DQB10301	827	374	0.813	**0.849**
HLA-DQA10201-DQB10303	761	265	0.743	**0.894**
HLA-DQA10201-DQB10402	768	241	0.529	**0.860**
HLA-DQA10301-DQB10301	207	66	0.822	**0.839**
HLA-DQA10303-DQB10402	567	117	0.483	**0.820**
HLA-DQA10501-DQB10302	847	203	0.772	**0.822**
HLA-DQA10501-DQB10303	564	179	0.809	**0.876**
HLA-DQA10501-DQB10402	749	337	0.584	**0.868**
HLA-DQA10601-DQB10402	565	133	0.498	**0.848**
Average			**0.719**	**0.826**

Table 4

			NetMHCIIpan-3.1-LOMO		NetMHCIIpan-3.2-LOMO	
Molecule	**#peptides**	**#binders**	**AUC**	**Pseudo distance 2013**	**AUC**	**Pseudo distance 2016**
DRB1_0101	2717	2599	0.735	0.22	**0.765**	0.16
DRB1_0301	1403	379	0.726	0.11	**0.739**	0.14
DRB1_0401	1639	695	0.757	0.04	**0.769**	0.04
DRB1_0404	542	331	0.765	0.06	**0.773**	0.03
DRB1_0405	1438	594	**0.822**	0.04	0.819	0.04
DRB1_0701	1619	806	**0.816**	0.28	0.814	0.27
DRB1_0802	1310	400	0.678	0.03	**0.700**	0.03
DRB1_0901	841	560	0.698	0.25	**0.713**	0.25
DRB1_1101	1604	730	0.720	0.06	**0.770**	0.06
DRB1_1302	1351	463	0.652	0.06	**0.662**	0.05
DRB1_1501	311	100	0.734	0.20	**0.802**	0.13
DRB3_0101	1266	267	0.685	0.12	**0.690**	0.14
DRB4_0101	1329	467	**0.752**	0.27	0.716	0.00
DRB5_0101	1606	765	0.802	0.20	**0.805**	0.20
H-2-IAb	39	20	0.818	0.34	**0.842**	0.34
H-2-IAd	107	24	0.778	0.34	**0.820**	0.34
HLA-DPA10103-DPB10201	5	1	1.000	0.06	1.000	0.06
HLA-DPA10103-DPB10401	1075	458	0.942	0.06	**0.952**	0.06
HLA-DPA10201-DPB10101	1180	558	**0.935**	0.07	0.927	0.07
HLA-DPA10201-DPB10501	1114	415	0.933	0.07	**0.938**	0.07
HLA-DPA10301-DPB10402	1193	498	0.934	0.09	0.934	0.11
HLA-DQA10101-DQB10501	990	246	**0.741**	0.23	0.687	0.02
HLA-DQA10102-DQB10602	1121	503	0.553	0.23	**0.786**	0.07
HLA-DQA10301-DQB10302	1461	330	**0.645**	0.19	0.619	0.09
HLA-DQA10401-DQB10402	1436	516	**0.875**	0.26	0.690	0.02
HLA-DQA10501-DQB10201	1386	477	0.547	0.27	**0.761**	0.07
HLA-DQA10501-DQB10301	1274	530	0.442	0.19	**0.647**	0.06
Average			**0.759**		**0.783**	

Table 5